Nyelvészeti tudásforrások integrálási lehetőségei diszkriminatív szegmens-alapú beszédfelismerő rendszerekbe

Tóth László,[1] Kocsor András,[2] Kovács Kornél,[3] Felföldi László [4]

MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport,
H-6720 Szeged, Aradi vértanúk tere 1., Hungary
{[1]tothl, [2]kocsor, [3]kkornel, [3]lfelfold}@inf.u-szeged.hu
http://www.inf.u-szeged.hu/speech

Kivonat A gépi beszédfelismerésben jelenleg kizárólag csak statisztikai elven működő algoritmusokat használnak. Ezek egyszerű matematikai modelleken alapulnak, amelyek paramétereiket hatalmas adatbázisokon automatikusan hangolják be. Az algoritmikai szempontok sajnos háttérbe szorítják a fonetikai/nyelvészeti ismereteket, így ezek a modellek irreális egyszerűsítő feltevéssekel élnek a beszédkommunikáció természetére nézve. Egy lehetséges alternatíva az ún. szegmentális modellek használata, amelyek – a statisztikai alapelv feladása nélkül – enyhébb megszorításokra épülnek. Ebben a cikkben bemutatjuk a tanszékünkön fejlesztett OASIS szegmens-alapú felismerőt, amely diszkriminatív elven, azaz posteriori valószínűségek összekombinálásával dolgozik. Ennek további előnye, hogy nagyobb rugalmasságot biztosít a különféle szintű (de továbbra is statisztikai jellegű) nyelvi információk integrálására, mint a hagyományos rejtett Markov-modell.

1. Bevezetés

"Mivel járult hozzá a fonetika a gépi beszédfelismeréshez?" - tette fel a kérdést nemrégiben az utóbbi terület egyik neves képviselője egy konferencián. A jelenlegi felismerők ugyanis nem fonetikai ismeretekre, hanem statisztikai elvekre épülnek. Ezek hatékonyságukat annak köszönhetik, hogy paramétereiket hatalmas méretű tanító adatbázisokon optimalizálhatjuk. Ennek ára, hogy ehhez viszonylag egyszerű (s így irreális feltételezéseken alapuló) matematikai modellt kell készítenünk. Ezért a jelenleg legelterjedtebb technológia, a rejtett Markov modellezés (HMM) [4] számos egyszerűsítő feltevéssel él a beszédjel információkódolási módjára nézve. A statisztikai megközelítés azonban nem zárja ki eredendően a fonetikai vagy percepciós ismeretek beépítését: ezeket a modell struktúrájának kialakításakor használhatjuk fel (ún. inductive bias). A HMM problémáinak egy részén túlléphetünk az általunk is használt ún. szegmentális modellekkel. Ezen túlmenően az általunk alkalmazott diszkriminatív modellezési technika egyszerű módot kínál a magasabb szintű (statisztikai jellegű) nyelvi tudásforrásoknak a felismerésbe történő integrálására. Cikkünkben bemutatjuk a tanszékünkön fejlesztett OASIS rendszer jelenlegi felépítését, majd megvizsgáljuk, hogy a diszkriminatív szegmens-alapú séma milyen lehetőségeket kínál a különböző szintű nyelvi modellek beépítésére.

2. Beszéd-dekódolás az OASIS rendszerben

A gépi beszédfelismerés célja egy A beszédjelhez a hozzá legjobban illeszkedő $F = f_1...f_N$ szimbólumsorozat megtalálása, ahol az f_n-ek beszédhangokat, vagy valamely más, alkalmasan megválasztott kódolási egységeket jelölnek. Jóformán minden felismerési algoritmus feltételezi, hogy mindegyik szimbólumnak megfeleltethető a jel egy jól meghatározott A_n időintervalluma, azaz a jel fonetikailag szegmentálható (ez az első, jelen esetben fonetikai jellegű nem teljesen reális feltevésünk a dekódolás során). Mivel a jelnek sem a helyes S szegmentálása, sem a valódi F fonetikai átirata nem ismert, ezért a felismerés során végig kell vizsgálni a jel összes lehetséges szegmentálását és az azokra illeszthető összes lehetséges szimbólumsorozatot. A feldolgozás során az összes lehetőséghez valamilyen költség-értéket rendelünk, és a legjobb költségű esetet fogjuk a felismerés eredményének tekinteni. Algoritmikailag az esetek végignézése egy keresési feladatot jelent, amelyet szaknyelven dekódolásnak nevezünk. Erre a célra az általunk fejlesztett OASIS rendszerben olyan algoritmust akartunk adni, amely kellően rugalmas ahhoz, hogy sokfajta stratégiát ki lehessen próbálni. Jelen pillanatban a rendszer az alábbi általános sémát használja.

Legyen az A jel valamilyen $a_1...a_T$ akusztikus események sorozataként adott. A keresés során balról jobbra haladva beszédhangokat igyekszünk ráilleszteni a mérésekre, minden lehetséges szegmenshatárt figyelembe véve. A művelet során megoldáskezdemények serege áll elő, ezeket hipotéziseknek fogjuk nevezni. Egy $H(t, F, w)$ hipotézis minimálisan a következőket tartalmazza: t az az időindex, ameddig az A feldolgozásában eljutottunk, $F = f_1...f_t$ a jel eddigi részére illesztett szimbólumsorozat, w pedig az illesztéshez rendelt költség. A dekódolás vázlatosan a következő: kiválasztunk egyet az eddigi hipotézisek közül, és kiterjesztjük. Ez abból áll, hogy az eddigi végponttól előre "tapogatózunk" valahány mérési adatnyit, és megvizsgáljuk, hogy ezek milyen eséllyel (költséggel) képezhetnek egy következő beszédhangot. Mivel nem tudjuk, hogy pontosan hány mérési adat tartozik a következő hanghoz, ezért minden lehetőségből egy-egy újabb hipotézist képezünk. Az új hipotézisek költségét úgy kapjuk meg, hogy az eddigi költséget és az újabb hang illesztésének költségét összekombináljuk egy megfelelő függvény segítségével. A túloldali pszeudo-kód részletezi az algoritmust.

A hipotézistér bejárására sokféle stratégia létezik, például az időszinkron bejárás, vagy a veremalapú dekódolás és variánsai [4]. Az algoritmus akkor ér véget, ha találtunk egy megoldást, vagy már nincs több kiterjeszthető hipotézis. Többnyire beérjük az előbbivel, azaz csak a legelső illeszkedő szimbólumsorozatot keressük meg. Megfelelően megválasztott bejárási stratégiával ugyanis garantálni (de legalábbis jó eséllyel biztosítani) lehet, hogy az első megoldás egyben optimális is. Az algoritmus esetleg úgy is véget érhet, hogy nem talál illeszkedő sorozatot. Ezt nagymértékben befolyásolja a vágási feltétel, amelynek segítségével elvethetjük az esélytelennek látszó hipotéziseket. Ezzel jelentősen csökkenthetjük a keresési teret, de esetleg a jó megoldás kidobását is kockáztathatjuk. A kiterjesztés új beszédhangjába bevont akusztikus események számát a megállási feltétel limitálja. Erre legegyszerűbb megoldás korlátot adni a beszédhangok lehetséges hosszára. De ha w_f az események számának monoton függvénye, akkor egy w_f-re adott küszöb is használható. Végezetül, a költségszámítást a g_1 és g_2 függvények végzik. Előbbi az egyes beszédhangok, utóbbi a teljes hipotézis költségét méri, és természetesen mindkettő kulcsfontossággal bír a helyes hipotézis megtalálásában.

Algorithm 1 Általánosított beszéd-dekódolási algoritmus

megoldáslista := $\emptyset$
hipotézislista := $h_0(t_0, "", 0)$
while van kiterjeszthető hipotézis **do**
  válasszunk egy $H(t, F, w)$ kiterjeszthető hipotézist valamely stratégia alapján
  if $t = T$ **then**
    if csak az első megoldás kell **then**
      return H
    else
      helyezzük át H-t a megoldások közé
    end if
  end if
  for $t' = t+1, t+2 \cdots$ **do**
    for all f **do**
      $w_f := g_1(f, < t, t' >)$ ahol g_1 az $f < t, t' >$-re való illeszkedési költsége
      $w' := g_2(w, w_f)$ ahol g_2 egy megfelelő aggregációs függvény
      if vágási-feltétel(w_f, w') **then**
        képezzünk egy új $H'(t', Ff, w')$ hipotézist és tegyük a hipotézislistába
      end if
    end for
    if megállási-feltétel($< t, t' >$) **then**
      break
    end if
  end for
end while

2.1. Speciális eset: Rejtett Markov-modell

A fenti dekódolási sémában elvileg sokféle módon lehetne a költségeket meghatározó g_1 és g_2 függvényeket megválasztani. A gyakorlatban azonban statisztikai módszereket használnak, azaz a költségek valószínűségi értékeknek felelnek meg. Ennek oka, hogy az ún. Bayes döntési elv alapján optimális működés (minimális számú tévesztés) garantálható [4]. Ehhez azonban szükséges egy jó becslés a $P(F|A)$ valószínűségre nézve. A gépi tanulás számos módszert ismer a valószínűségek példák (tanító adatbázisok) alapján történő becslésére, a beszédfelismerési probléma azonban speciális, amennyiben a lehetséges A beszédjelek és F átiratok száma potenciálisan végtelen. Erre az egyetlen megoldás mindkettőt kisebb egységekre bontani. Ekkor kerül képbe a már említett szegmentális feltevés, azaz a jelet szegmentumokra bontjuk, és $P(F|A)$-t a szegmentumokhoz rendelt $P(f_n|A_n)$ értékekből kombináljuk össze. Erre a valószínűségszámítás lényegében egyetlen egyszerű módot kínál: ha az egységek függetlenek, akkor a valószínűségek összeszorozhatók. Ez a másik – ez esetben matematikai jellegű – olyan egyszerűsítő feltevés, amellyel élni fogunk, habár valójában nem teljesül.

Dekódolási sémánkba belefér a rejtett Markov-modellezésnek az az esete, amikor a modell szigorúan balról-jobbra típusú, és állapotok átugrása nem lehetséges. Folytonos beszéd felismerésére legtöbbször ilyet szoktak használni, három állapottal, ahol az állapotok (nálunk az f_n-ek) beszédhang-harmadoknak (kezdő-átmeneti, stabil, záró-átmeneti szakasz) felelnek meg [4]. A dekódolás során tehát beszédhangok helyett eze-

ket kell használnunk fonetikai szimbólumként. Az akusztikai megfigyeléseket 10-30 ezredmásodpercenként (szakszóval "keretenként") számolt a_l spektrális adatok képezik. A g_2 aggregációs függvény egyszerűen csak összeszorozza az egyes szimbólumokhoz rendelt értékeket. A lényeg a g_1 függvény, amely az alábbi alakot ölti:

$$g1(f, < t, t' >) = P(f|F) \cdot l_f^{(t'-t)} \cdot \prod_{i=t}^{t'} P(a_i|f), \quad (1)$$

ahol a harmadik tényező a $< t, t' >$ intervallum minden a_l méréséhez kiszámol egy $P(a_l|f)$ valószínűséget, és – függetlenséget feltételezve – ezeket összeszorozza. A második tényező egy exponenciálisan lecsengő hosszmodell, amely egy megfelelően beállított l_f konstanst igényel. Végezetül $P(f|F)$ a nyelvi modell hozzájárulása. A második két tényező ilyen módon való felírásának technikai okai vannak: mivel a különböző $< t, t' >$ intervallumokhoz rendelt költségek egymást tartalmazzák, így megfelelő technikával (dinamikus programozás) párhuzamosan, s így gyorsan számíthatóak. Egy további, talán még fontosabb szempont, hogy a modell komponenseit képező $P(a_l|f)$ eloszlások és l_f konstansok adatbázisok alapján történő tanulására hatékony algoritmusokat lehet adni (a $P(f|F)$ nyelvi modellt ezektől függetlenül szokás tanítani).

2.2. Speciális eset: Szegmentális modellek

Az adott beszédhanghoz sorolt $P(a_l|f)$ értékek összeszorzása hatékony ugyan, de a függetlenségi feltevés erősen irreálisnak tűnik. A hosszt leíró komponens exponenciális lecsengése sem felel meg a gyakorlati méréseknek. Ezen problémák feloldására javasolták az ún. szegmentális modellek használatát, amelyek a $< t, t' >$ szegmenst „egyben" modellezik [9]. Ennek lényege, hogy a HMM-nél látott $g_1(f, < t, t' >)$ számítási képlet második két tényezőjét lecseréljük valamilyen műveletigényesebb, de az adatoknak remélhetőleg jobban megfelelő modellre. A bonyolultabb megoldások parametrikus modelleket illesztenek a $< t, t' >$-hez tartozó a_l adatokra [9]. Az egyszerűbb megoldás először is $< t, t' >$-t annak hosszától függetlenül ugyanannyi adattal próbálja meg leírni. Ennek legkönnyebb módja az adatokat elsimítani, és fix számú mintát venni belőle [3]. Értelme pedig az, hogy az így kapott reprezentáción immár alkalmazható az a rengeteg fajta modellezési technika, amelyet a gépi tanulásban valaha felvetettek rögzített dimenziószámú terekben való osztályozásra, illetve valószínűségi regresszióra.

Egy további szempont is felvetődik itt, mégpedig az, hogy a rejtett Markov modell az ún. generatív modellek családjába tartozik. Ez azt jelenti, hogy a $P(a_l|f)$ valószínűségekből építkezik, szemben az ún. diszkriminatív modellekkel, amelyek az f szimbólumok $P(f|a_l)$ a posteriori valószínűségével dolgoznak. Ennek két okból van jelentősége: az egyik, hogy tapasztalatok szerint a diszkriminatív modellek kicsit jobb osztályozási eredményeket képesek elérni (bonyolultabb tanítási folyamat árán), mint a generatívak. A másik, hogy amennyiben a felismerés során többféle tudásforrást akarunk kombinálni, akkor erre a diszkriminatív modellezés sokkal többféle módot és lehetőséget kínál.

Az OASIS rendszerben az utóbbi években számos algoritmust kipróbáltunk beszédhangok diszkriminatív szegmens-alapú osztályozására. Az irodalommal összhangban mi is azt találtuk, hogy ez az egyszerű séma valamivel jobb beszédhang-felismerési eredményekre képes, mint a keret-alapú modellezés [7] [8] [11].

3. Nyelvi modellezés az OASIS jelenlegi verziójában

Mint láthattuk, a jelenlegi beszédfelismerési technika alapvetően valószínűségi alapú, és a nyelvi modelltől is azt várja, hogy valószínűségeket rendeljen a nyelvi egységekhez. Természetesen a hagyományos szabályalapú nyelvleírások is tekinthetők ilyennek, hiszen felfoghatók úgy, mintha kizárólag 0 és 1 valószínűségeket adnának ki. A gyakorlatban az ilyen modellek azonban túl merevnek bizonyultak, így érdemesebb az engedékenyebb valószínűségi modellekhez fordulni. Szerencsére a szabályalapú technikák közül több kiterjeszthető valószínűségi jellegűvé, így kaphatjuk például a sztochasztikus környezetfüggetlen nyelvtanokat (P-CFG) vagy a súlyozott automatákat.

Az OASIS rendszer nyelvi moduljának megtervezésekor igyekeztünk követni a más felismerőkben definiált nyelvleírási technikákat. Ehhez a Microsoft Speech API-ból indultunk ki, amelyben környezetfüggetlen nyelvtanokat lehet definiálni egy XML leírási formátumot követve. Mivel a SAPI-t angol nyelvre találták ki, így a nyelvtanok nemterminálisai közvetlenül a nyelv szavai. A magyar nyelv esetében viszont az összes lehetséges toldalékolt alak felsorolása kezelhetetlen. Szerencsére a magyar morfológia modellezése véges állapotú automatákkal jól megoldható [2]. Tapasztalatunk szerint egy adott szó toldalékolt alakjai automatával nagyságrendekkel kisebb helyen tárolhatóak, mint bármilyen hagyományos tömörítőprogrammal. Ezért a SAPI leírást kiterjesztettük oly módon, hogy nálunk a terminálisok helyére automatákat is be lehet ágyazni. Ez egy olyan környezetfüggetlen nyelvtanhoz vezet, amelynek terminális szimbólumai az automaták által felismert nyelv szavai. További tömörítést érhetünk el a morfológiát kezelő automaták tömör reprezentációjával. Ehhez speciális automatatömörítő algoritmusokat használunk, amelyek az adott nyelvet felismerő automaták közül a lehető legkisebbet konstruálják meg [5]. További tárcsökkentést jelent, ha az automatát is kis helyigényű adatszerkezettel, például a [6]-ban megadott módszer szerint tároljuk.

A valószínűségek kezelésére a Speech API-ban súlyokat rendelhetünk a szabályok alternatív jobboldalaihoz. Az OASIS nyelvi moduljában kiegészítésként bevezetett automaták szintén megengedik az egyes elágazások súlyozását, így a két szint kombinálásával a rendszer képes az egyes beszédhang-sorozatokhoz valószínűségeket rendelni.

A nyelvi modell interfészének kialakításánál figyelembe kellett vennünk, hogy a felismerést végző (azaz az 1. algoritmust végrehajtó) modul milyen formában várja a nyelvi modell támogatását. Mivel a hipotézisek kiterjesztése során egy adott hangsorozat lehetséges folytatásaira van szükségünk, ezért a nyelvi modul feladata egy adott prefix összes lehetséges folytatásait (azaz a következő beszédhangot) visszaadni. Így a nyelvi modul interfésze az alábbi két függvényből áll, amelyek iterátor-jellegű bejárást biztosítanak a nyelvi modell összes lehetséges beszédhang-sorozatának végignézéséhez:

`Enter`: A megadott prefixhez meghatározza az első lehetséges kiegészítést, és annak valószínűségét (ha nincs ilyen, akkor null-t ad vissza).

`Next`: Megadja (az ugyanazon prefixhez tartozó) következő lehetséges kiegészítést, és annak valószínűségét. Ha nincs több, akkor null-t ad vissza.

Technikai szempontból a modell automatáinak implementálása illetve bejárása viszonylag egyszerűen megoldható. A környezetfüggetlen nyelvtan kezelése azonban már veremautomata használatát igényli. Ebben az esetben a verem aktuális értékeinek tárolása is szükséges, ami a nyelvi modul megvalósítását megnehezíti.

4. További lehetséges nyelvészeti tudásforrások integrálása

Az 1. algoritmusban leírt dekódolási séma kellően általános, így könnyen kiterjeszthető nagyobb számú információforrás egyesítésére. Ehhez csak a g_1 (és esetleg a g_2) függvény(eke)t kell megfelelően módosítani. Gyakorlati szempontból a legkritikusabb pont, hogy a modellt ne bonyolítsuk el annyira, hogy az optimális paraméterek algoritmikus megtalálása lehetetlenné váljék. Szerencsére sok olyan matematikai módszer ismert, amely tudásforrások optimális integrálásáról szól, illetve az osztályozók kombinálása is rendkívül aktív kutatási téma az utóbbi időben. Továbbá a beszédfelismerésben egyre inkább terjednek az olyan optimalizálási módszerek, amelyek az általunk is használt diszkriminatív modellezést támogatják [10]. Ilyen például az ún. diszkriminatív modell-kombinálási technika, mellyel az alábbi típusú integrálást optimalizálhatjuk [1]:

$$P(F|A, L_1, ..., L_r) \approx \max_S \prod_i P(f_i|A, S)^{\alpha_0} P(f_i|L_1)^{\alpha_1} \cdots P(f_i|L_r)^{\alpha_r}, \quad (2)$$

ahol ez esetben r darab (nyelvi) információforrásunk van, $L_1, ..., L_r$, és ezek posterior valószínűségek formájában „szavaznak" az egyes f_i szimbólumokra. A források kombinálása hatványozás, majd szorzás útján történik. Természetesen másfajta kombinációval is próbálkozhatunk, de az optimális kombinálás megtalálása más esetekben más matematikai elveket kívánhat. Az OASIS rendszerben jelen állapotban még csak egyetlen nyelvi modell van (az előző fejezetben leírtaknak megfelelően), és kombinálási szabályként a hagyományos rendszerekben már bevált szorzást alkalmazzuk, azonban többfajta alternatív kombinálási technika kipróbálását is tervezzük a közeljövőben.

Hivatkozások

1. P. Beyerlein, Discriminative Model Combination, Proc. ICASSP'98, pp. 481-484., 1998.
2. Futó Iván (szerk.), Mesterséges intelligencia, Aula, 1999.
3. J. R. Glass, A probabilistic framework for feature-based speech recognition, Proc. ICSLP'96, pp. 2277-2280, 1996.
4. X. D. Huang, A. Acero és H-W. Hon, Spoken language processing, Prentice Hall, 2001.
5. Kertész-Farkas Attila, Fülöp Zoltán, Kocsor András: Magyar nyelvű szótárak tömör reprezentációja nemdeterminisztikus automatákkal, Ugyanebben a kiadványban
6. G. A. Kiraz, Compressed Storage of Sparse Finite-State Transducers, *Proc. of WIA'99 (Szerk. O. Boldt és H. Jürgensen)*, LNCS Vol. 2214, pp. 109-122, Springer, 2001.
7. Kocsor A. et al., A Comparative Study of Several Feature Space Transformation and Learning Methods for Phoneme Classification, Int. J. Speech Technology, Vol. 3, 3/4, pp. 263-276, 2000.
8. Kocsor, A. és Toth, L., Application of Kernel-Based Feature Space Transformations and Learning Methods to Phoneme Classification, elfogadva az Applied Intelligence folyóiratba
9. M. Ostendorf, V. Digalakis és O. A. Kimball, From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, IEEE Trans. ASSP, 4:360–378., 1996.
10. R. Schlüter et al., Comparison of discriminative training criteria and optimization methods for speech recognition, Speech Communication, Vol. 34., pp. 287-310., 2001.
11. Tóth L., Kocsor A. és Kovács K.: A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition, Proceedings of TSD'2000 szerk. P. Sojka, I. Kopecek és K. Pala, LNAI 1902, pp. 307-313, Springer Verlag, 2000.